

Network Classification and Categorization

James P. Canning², Emma E. Ingram³, Sammantha Nowak-Wolff¹,
 Adriana M. Ortiz⁴, Nesreen K. Ahmed⁵, Ryan A. Rossi⁶,
 Karl R. B. Schmitt¹, and Sucheta Soundarajan⁷

¹ Valparaiso University, Valparaiso IN 46383, USA,
 Corresponding Author: karl.schmitt@valpo.edu

² SUNY Geneseo, Geneseo NY 14454, USA,

³ University of Alabama, Tuscaloosa AL 35487, USA,

⁴ University of Puerto Rico, Rio Piedras, San Juan PR 00936,

⁵ Intel Labs, 3065 Bowers Ave, Santa Clara, CA USA

⁶ Xerox PARC, 3333 Coyote Hill Rd, Palo Alto, CA USA

⁷ Syracuse University, 223 Link Hall, Syracuse, NY USA

1 Introduction

Networks are often categorized according to the underlying phenomena that they represent, such as re-tweets, protein interactions, or web page links. It is generally believed that networks from different categories have inherently unique network characteristics. In this work, we find strong evidence supporting this hypothesis by learning a classification model $f : \mathbf{x} \rightarrow y$ that is able to *accurately* predict (with 94.2% accuracy) the category of a new arbitrary unknown network G' described only by a D -dimensional feature vector \mathbf{x}' where $y \in \{1, 2, \dots, K\}$ is the class label (category). The classifier f is learned using over $N=500$ networks from $K=8$ categories (See Figure 2) which are characterized using only $D=15$ simple structural features (Table 1). As an aside, Graphlet features [1] and other more discriminative features can be used to further improve the accuracy.

To the best of our knowledge, this work is the first large-scale study that tests whether network categories are distinguishable from one another (using both categories of real-world networks and synthetic graphs). Previous research has focused on either (i) classification of synthetic graphs or (ii) graphs within a particular category such as molecular graphs. Other examples include distinguishing between brain or breast cancer cells [2] or distinguishing between different social structures [3].

A classification accuracy of 94.2% was achieved using a random forest classifier with both real and synthetic networks. These results indicate that while some of the categories researchers use to label their graphs are indeed distinct, others, from a feature standpoint, are largely indistinguishable from one another. Moreover, from a feature standpoint, synthetic graphs are trivial to classify as they are structurally distinct from all other graphs. Additionally, the classifiers also highlighted networks that are outliers within their own categories, suggesting new potential directions for understanding those networks.

2 Data

Data was originally pulled from the Network Repository [4] for all non-synthetic graphs. This included 1241 graphs with 15 network features. The features in

the data are listed in Table 1. Of the 20 network categories included, three were from computational and algorithmic challenges (DIMACS, DIMACS10 and BHOSLIB) and two recorded graphs over time (Temporal Reachability, Dynamic Networks). As all five of these categories are fundamentally different from static one-time recorded networks from a discipline or field they were discarded as outside the problem scope. Within the 15 remaining categories, 9 categories had less than 20 instances and thus were also excluded as having insufficient data for training. Finally, Cheminformatics had significantly more instances than all other categories and therefore was downsampled to 119 networks which is comparable to the 2nd largest category. We also generated 125 graphs: 50 using the Barabasi-Albert (BA) model and 75 using the Erdős-Rényi (ER) model. The final classification data set has 529 graphs from 8 categories.

Number of Nodes	Avg. Degree	Avg. Clustering Coefficient	Assortativity
Number of Edges	Min. Degree	Fraction of Closed Triangles	Total Triangles
Maximum K-core	Max. Degree	Max. Clique (lower bound)	Avg. Triangles
Chromatic Num.	Density	Maximum Triangles	

Table 1. Features calculated by the Network Repository

3 Results

Evidence from both unsupervised and supervised machine learning (ML) algorithms points to clear, distinctive structure in real-world networks from different domains. Dimensionality reduction using t-distributed stochastic neighbor embedding (t-SNE) shows clear clusters of graphs (see Figure 1. Specifically, Facebook, Cheminformatics, Retweet, Brain, Social and Web/Technological graphs are able to be identified both visually and using k-means clustering.

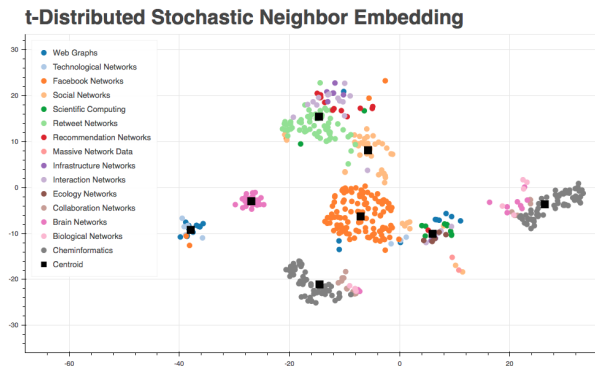


Fig. 1. t-SNE Clustering. Black Squares indicate centroids from K-means

Similarly, standard classification algorithms are able to accurately classify graphs from each of those categories. A summary of the classification results are shown in Figure 2 and supports several important findings. First, we see that even though Erdős-Rényi (ER) and Barabasi graphs (BA) are intended to model real-networks, they are distinct enough from their inspirations that only two other networks are classified as either BA or ER. This result questions

		PREDICTED								Recall
		B	C	F	R	S	W	BA	ER	
ACTUAL	Brain	31	0	1	2	1	0	1	0	0.86
	Chem	0	119	0	0	0	0	0	0	1
	Facebook	0	0	113	1	0	0	0	0	0.99
	Retweet	0	0	0	59	3	0	0	0	0.95
	Social	0	0	1	1	40	5	0	1	0.83
	Web	1	0	0	1	10	10	0	0	0.45
	Barabasi	0	0	0	0	0	0	50	0	1
	Erdős-Rényi	0	0	0	0	0	0	0	75	1
	Precision	0.97	1	0.98	0.92	0.74	0.67	0.98	0.99	

Fig. 2. Contingency Matrix for Classification from a Random Forest Model

the efficacy of testing algorithms/ideas intended for real networks on synthetic models. Second, it is apparent that graphs normally labeled “Web” are difficult to distinguish from social graphs. Deeper evaluation reveals that several of the web graphs represent pages within a specific social community, which could therefore influence the network’s structure. Likewise, several social graphs are from very techno-centric realms and could be reasonably labeled as a web graph.

Additional tests show that two categories of graphs initially excluded due to low instances could be combined with existing categories with a minimal loss in accuracy. Results from k-means clustering indicate that Web and Technological graphs as well as Brain and Biological networks have similar properties. Finally, careful analysis of the mislabeled graphs in Figure 2 provides interesting network/category specific findings and suggestions. For example, 10 of the 36 brain networks are non-human, however all 5 graphs that are mislabeled are non-human. This is strong evidence that either the human networks are truly distinct from the non-humans, or the network discovery process is not sufficiently standardized for neuro-networks. Also interesting was that a visual inspection of the graph mislabeled as a retweet network shows surprising similarities. This suggests that using classification models provide valuable insight into alternative research techniques for crossing disciplines.

4 Conclusions

This work makes two important findings. First, real-world networks from various domains have distinct structural properties that allow us to predict with high accuracy the category of an arbitrary network. Second, classifying synthetic networks is trivial as our models can easily distinguish between synthetic graphs and the real-world networks they are supposed to model.

References

1. Ahmed, N.K., Neville, J., Rossi, R.A., Duffield, N.: Efficient graphlet counting for large networks. In: ICDM. (2015) 1–10
2. Li, G., Semerci, M., Yener, B., Zaki, M.J.: Effective graph classification based on topological and label attributes. *Stat. Anal. and Data Mining* **5**(4) (2012) 265–283
3. Ugander, J., Backstrom, L., Kleinberg, J.: Subgraph frequencies: Mapping the empirical and extremal geography of large graph collections. In: WWW. (2013) 1307
4. Rossi, R.A., Ahmed, N.K.: The network data repository with interactive graph analytics and visualization. In: AAAI. (2015)